Reading Comprehension

- **Understanding as per our discussion:**
  - Question generation aims to generate questions from a text passage where the generated questions can be answered by certain sub-spans of the given passage.
  - Traditional methods mainly use rigid heuristic rules to transform a sentence into related questions. In this work, we will use neural encoder-decoder model to generate meaningful and diverse questions from natural language sentences as our final project.

- Some more research done by us :
  - Similarity Approach Baseline - It uses word embedding and building triples (S,V,O). It works by finding closest sentence from passage and then get the missing word or phrase and output it.
  - Using Framenet Framework - It includes running FrameNet both on the passage and the question. Find similar frames, Do matching/rules to retrieve the answer.
    Both of these techniques are used to give answers from passage for given Questions.

- **Baseline we chose to implement** :

  <u>Question Generation:</u>

  - Each Passage is broken down into sentences.
  - Each sentence is parsed using English grammar rules with the use of condition statements.
  - A dictionary is created called **bucket** and the part-of-speech tags are added to it.The sentence which gets parsed successfully generates a question sentence.
  - The generated question list is printed as output.
  - We can use a dataset of text and questions along with machine learning to ask better questions. (Not included in baseline)

Tasks to be done in initial phase :

- Completing the Baseline implementation.
- Understanding the outputs from baseline and finding out the ways to improve these outputs.
- Reading about machine learning techniques which can be used on top of the baseline to get better results.

- **Datasets considered** :

  Some available datasets which we searched for the project :
  - Microsoft MARCO Dataset: 100,000 English queries along with corresponding passages and answers. This is a great data set because it contains real questions asked by humans on Microsoft search engines. Microsoft also provided more than one passages which might be able to answer the question. If our model can work well on this data set which means we are close to providing answers to the real search engines.
  - Stanford Question Answering Dataset (SQuAD): 100,000 English question-answer pairs on 500 articles . SQuAD consists of questions posed by crowd workers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage. This dataset fit the definition of our task perfectly.

  We will be using **SQuAD**

  Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.

  (the implementation aspects may change as we shall proceed further with new ideas.)